Lab 7: F-test and dummy variables

Statistics for Social Scientists II

Bur, GJM

2024-10-21

1 The F-test simplified

I want to start by noting an important way of understanding our F-test. Note that it can always be written as a simple ratio. *In general, it is always* $\frac{m(\hat{\mathbf{y}})}{m(\hat{\mathbf{e}})}$, where $m(\hat{\mathbf{y}})$ is the mean model sum of squares for some model whose significance we want to test and $m(\hat{\mathbf{e}})$ is the mean (in-sample) error sum of squares or the residual sum of squares for some comparison model.

How does this relate to our familiar formula? Well, in the case of the *model* F-test, we have …

$$\frac{SS_e^r - SS_e^f}{df^r - df^f} \div \frac{SS_e^f}{df^f}$$

Since $SS_t - SS_e = SS_m$, and our reduced model is just the mean-only model, hence $SS_e^r = SS_t^f$, the numerator is just SS_m^f. Then, the denominator in the first term is just the degrees of freedom for the regression space of the full model. So, the first term is just the mean model sum of squares; the denominator is just the mean square error for our model.

What about in the general case?

Let's now write the total sum of squares like so: $SS_t = SS_e^f + SS_m^r + SS_m^{r\perp}$. What this means is that the total sum of squares in the full model can be decomposed into the sum of errors for the full model plus the model sum of squares *with* our restrictions in place (i.e., the model sum for the reduced model) plus the model sum of squares that the full model adds, the incremental sum of squares $SS_m^{r\perp}$. The "perpendicular" sign refers to the fact that in subject space, the regression space is decomposable into two perpendicular or orthogonal parts, one with our restriction in place and the other which captures remaining regression effects beyond that restriction.

Then, $SS_t - SS_e^f - SS_m^r = SS_m^{r\perp}$. Since $SS_t - SS_m^r = SS_e^r$, we have that $SS_e^r - SS_e^f = SS_m^{r\perp}$. So our numerator measures very simply the incremental model sum of squares gained by adding the incremental predictors to the model; the denominator is the error sum of squares for that full model. Intuitively, if the gain from adding those predictors is big relative to the total error, they are significant!

2 Review: dummy variables generally, simple regression on them

In what follows, all references to dummy variables assume that "dummy" means a variable D which takes values \$ d $\{0, 1\}$\$. There are other possible meanings for "dummy variable", the most common of which is an "effects-coded" dummy variable D which takes values \$ d $\{-1, 0, 1\}$\$. We won't cover those here, but see appendix if you want.

Recall that at the level of the population, $\mu_D = p$, where p is the probability of success on the dummy, and the variance is $\sigma_D^2 = p(1-p)$. In the sample, the former holds exactly: $\bar{D} = \hat{p}$, where $\hat{p}$ is the sample probability. The variance is slightly different; it is $\frac{n}{n-1}\hat{p}(1-\hat{p})$.

2.1 Some hands-on practice: making dummy variables in Stata

Make a dummy variable for someone being married using `marstat` on the CPS 2019 extract I've posted before (full data here), using the following techniques.

1. The `gen if` approach (make a value label).
2. The Boolean assignment (make a value label).
3. The `recode` approach (this makes a value label).
4. The `tab var, gen(dummies)` approach (bonus, not mentioned by Gordon).
5. (Regression only): the factor variables syntax, `reg y i.polytomousvar x_2 ...`

2.2 Regression on a single dummy

We know that regression on a single dummy, i.e. estimating the model (in the population or sample) $Y = \beta_0 + \beta_1 D + \epsilon, D \in \{0, 1\}$, returns $\mu_{Y|D=1} - \mu_{Y|D=0}$: in words, the difference in the conditional means for the groups.

Here's a quick proof for the sample.

$$
\begin{aligned}
\beta_1 &= \frac{s_{X,Y}}{s_X^2} \\
&= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \\
&= \frac{p\bar{y}_1 - p\bar{y}}{p(1-p)} \\
&= \frac{\bar{y}_1 - \bar{y}}{(1-p)} \\
&= \frac{\bar{y}_1 - p\bar{y}_1 - (1-p)\bar{y}_0}{(1-p)} \\
&= \frac{\bar{y}_1(1-p) - (1-p)\bar{y}_0}{(1-p)} \\
&= \bar{y}_1 - \bar{y}_0
\end{aligned}
$$

Try this out using your dummy for whether someone is married and the CPS

`wage4` variable, which is wages for all respondents regardless of job type. Cross-check your work with `tab $categorical$, sum($continuous$)`.

3 Regression on a set of dummies encoding one polytomous variable

We can also construct a model where we have many dummies representing a polytomous variable, i.e. one with many possible values. Then, we can construct a model $Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 ... + \epsilon = \beta_0 + \sum_{k=1}^{m} \beta_k D_k + \epsilon$. However, before we talk about what the coefficients in this model mean, we need to say something about the dummy variable trap.

3.1 The dummy variable trap

A common problem—one that some of you have already run into—is that of the "dummy variable trap", AKA multicollinearity. Simply put, this happens when one of the variables in your set of regressors is a perfect function of one of the others *or a linear combination of them.* It is also typically a problem for this to be *nearly* true, which we call "near-multicollinearity".

If you attempt to use all of the dummy variables corresponding to a single polytomous variable in a regression, the resulting mechanical problem is the dummy variable trap. The idea is pretty straightforward. In *subject space*, you can imagine this easily: one of our predictor vectors is either a scalar multiple of one of the existing vectors or it lies in the hyperplane formed by the others. So, one of the vectors isn't necessary to find a least-squares solution, and it's actually not clear what role it should play, so our regular OLS algorithm fails.

In algebraic terms, it's pretty straightforward as well. Consider a set of dummy variables made from, say, `marital` : `married`, `nevermarried`, `divorced`, `separated`, and `widowed`. Then, any of the variables we have is just the linear combination of the others. For example,

$$M = -1(N + D + S + W) + 1$$

In short, this basically makes our model impossible to uniquely estimate. If you want a small picture into how the underlying algebra of regression works, consider the following system of m equations, where m is the number of categories of our dummy variable. Then, it turns out that to estimate our m dummy coefficients plus one intercept, we need the following system of $m + 1$ equations in $m + 1$ variables.

$$\begin{bmatrix} \hat{\beta}_0 n + \sum_{k=1}^{m} \hat{\beta}_d n_k \\ \hat{\beta}_0 n_1 + \hat{\beta}_1 n_1 \\ \hat{\beta}_0 n_2 + \hat{\beta}_2 n_2 \\ \vdots \\ \hat{\beta}_0 n_m + \hat{\beta}_m n_m \end{bmatrix}$$

$$\begin{bmatrix} n\bar{y} \\ n_1\bar{y}_1 \\ n_2\bar{y}_2 \\ \vdots \\ n_m\bar{y}_m \end{bmatrix}$$

Unfortunately, it turns out that this system of $m+1$ equations in $m+1$ variables is technically not solvable because the sum of the bottom m equations is equal to the first equation. We essentially have $m+1$ unknowns but only m equations.

If we treat this as a linear algebra problem to be solved *directly*, the solution is to set a constraint, and one common constraint is to require that $\beta_1 = 0$. Then, the second equation above immediately requires that $\hat{\beta}_0 = \bar{y}_1$, and it follows that every $\beta_d = \bar{y}_d - \bar{y}_1 : d \neq 1$. *This turns out to be equivalent to simply omitting the variable X_1 from the model* and solving the following system of equations.

$$\begin{bmatrix} \hat{\beta}_0 n + \sum_{k=2}^{m} \hat{\beta}_k n_k \\ \hat{\beta}_0 n_2 + \hat{\beta}_2 n_2 \\ \vdots \\ \hat{\beta}_0 n_m + \hat{\beta}_m n_m \end{bmatrix} \begin{bmatrix} n\bar{y} \\ n_2\bar{y}_2 \\ \vdots \\ n_m\bar{y}_m \end{bmatrix}$$

3.2 Interpreting coefficients on dummies encoding a single polytomous variable

Then, it's easy but tedious to show that if we follow the above approach, coding a set of $m-1$ dummy variables as simple 0/1 binary variables and including them all in our model, ***each of the dummy coefficients expresses the difference in means between the reference group, i.e. the group which we exclude from the model, and the group whose coefficient it is***.

Try this out with the `marstat` variable.

You should try this two ways: first, make a set of dummies, try to cause the dummy variable trap to convince yourself it is real, and then try excluding one category. Then, try the syntax `reg $continuous$ i.$yourpolytomousvar$`. You can then check your work with `tab $cat$, sum($cont$)` by running `margins i.$yourpolytomousvar$` right after the regression.

4 Multiple sets of dummies

Why would we include multiple sets of dummies? From a modelling perspective, it would simply be because we think that the model with two sets *true*. From a more general predictive perspective, it's because we might think that by including some second set of variables, our model removes some spurious associations. This provides a nice way to interpret a model with multiple sets of dummies. The effect of *one* dummy on some polytomous variable P_1, controlling for another set of dummies representing P_2, represents the effect of switching from one category on P_1 to another, controlling for the effect that P_2 has on grouping within levels of P_1

Here is a stylized example in theory and with code. Suppose that P_2 is someone's aptitude measured in five levels; suppose that P_1 is their degree category. Suppose that academic achievement reflects aptitude, that aptitude is more closely related to earnings than schooling, and that schooling's other determinants are factors unrelated to earnings potential. Let Y_1 be earnings. Then, regressing earnings on degree category will produce significant results since someone's aptitude category strongly influences their degree category, but once we account for aptitude, it becomes insignificant.

```
set seed 4301963
clear all
set obs 1000
gen p_2 = runiformint(1, 3)
gen p_1 = p_2 + runiformint(0, 4)
gen y_1 = p_2 + runiform(0, 0.25)
reg y_1 i.p_1
reg y_1 i.p_1 i.p_2
```

4.1 Dummies and continuous variables

How do we interpret dummy regression when we include other ratio predictors, e.g. $Y = \beta_0 + \beta_1 X_1 + \beta_2 D + \epsilon$? It is the same sort of story. Suppose that our continuous variable is something like underlying IQ, to modify the previous example. Then, changes in educational status *net of* IQ would be essentially irrelevant to income.

```
gen cont = p_2 + rnormal(0, 1)
gen y_2 = cont + rnormal(0, 0.25)
reg y_2 i.p_1
reg y_2 i.p_1 cont
```

5 Three ways to find a difference in included means

Finally, how do we find the difference in means between *included* groups? There are three main ways Gordon mentions.

5.1 Re-estimate the regression with one as the reference group

This is pretty easy to do. Simply pick a different excluded group so that one of them corresponds to the difference which you want to test. You can do this very easily in Stata with `reg $cont$ ib$k$.$categorical$` where `k` refers to some value of the dummy variable that we want to serve as our base.

Try making divorced people the excluded group for our ongoing regression of wages on marriage; evaluate the difference between them and the never-married.

5.2 Partial F-test

We can also run a partial F-test using the restriction that our two slopes are the same. You can try doing this by hand and then with Stata's `test` command, which I should note is only

technically an approximation to the F-test (it calculates what is called the Wald statistic, but these converge).

Here is an example for you since this one is a bit more complicated.

```
clear all
use ./data/cps2019extract
tab marstat, gen(mardum)
fre marstat
recode marstat (2 3 = 1 "div. or wid.") (1 4 5 = 0 "not div. or wid.") ///
    (.=.), gen(divwid)
reg wage4 divwid mardum4 mardum5
local SSE_r = e(rss)
reg wage4 mardum2 mardum3 mardum4 mardum5
local SSE_f = e(rss)
di `SSE_f'
local df_ef = e(df_r)
di `df_ef'
local incr_F = (`SSE_r' - `SSE_f') * (`df_ef'/`SSE_f')
di `incr_F'
qui reg wage4 mardum2 mardum3 mardum4 mardum5
test mardum2 == mardum3
```

5.3 Post-regression test of a linear combination

Finally, we could use a linear combination of our estimated regression slopes. The idea here is that these random variables have standard errors just like any other random variables. In general, we know that the formula for a sum of random variables (and a difference is just a sum of a negative and positive variable) is given (from last week) by …

$$Y := Y_1 + Y_2 + ...Y_k = \sum_{p=1}^{k} Y_k$$

$$\sigma_Y^2 = \sum_{p=1}^{k} \sigma_p^2 + 2 \sum_{q>p}^{n} \sum_{p=1}^{n} \sigma_{p,q}$$

So here we can find the standard error of the difference $\hat{\beta}_p + -\hat{\beta}_q$ by summing their individual variances and then subtracting twice their covariance. This statistic happens to be t-distributed.

```
qui reg wage4 mardum2 mardum3 mardum4 mardum5
local diff = r(table)[1,1] - r(table)[1,2]
estat vce
local vardiff = r(V)[1,1] + r(V)[2,2] - 2*r(V)[1,2]
di `diff'/sqrt(`vardiff')
lincom mardum2 - mardum3
```